TTIC 31050: Assignment 2

Graham Northrup

January 30, 2018

0.1 Introduction

This assignment revolves around using using blocks of ungapped alignments to create the scoring matrices using the same techniques to make the BLOSUM matrices. Because my file input-output handling is really bad, I have hand cut a number of subsequences into my file. For the super family data set I used subsequences that all align to the first 43 Amino Acids of d1a6m__ in group 1. This gives me the ungapped alignment necessary to proceed.

For the twilight dataset, I am using a set of sequences from group 180. I use more in this group in an attempt to get more predictive power for across a more separated family of sequences.

It is worth pointing out here that I am running this algorithm on a much smaller set of blocks than BLOSUM, which will almost certainly impact my final matrix outside what we would see otherwise.

0.2 Implementation

I wrote a function called make_mat, which takes the list of sequences and the length of the block and returns the scoring matrix. This function works exactly as we described in class. One important implementation difference is the initialization of the mutation count matrix with ones instead of zeros. This is not exactly correct, but prevents the log calculation from breaking when $p_{ij} = 0$ due to the small sample size.

Otherwise, the function creates and returns the scoring matrices. I also ended up making a symmetrical scoring matrix across the diagonal instead of saving a .csv file half full of empty entries.

0.3 Comparison

The matrices generated by my code have been inserted into the report for ease of access. The matrices appear similar in several key ways including a number of favorable scores along

Table 1: Superfamily Matrix

A	1.12	-2.19	-0.09	-0.23	-0.03	1.37	0.43	0.49	-0.19	-0.57	-1.48	1.00	-2.59	-2.48	-0.81	0.41	0.45	0.42	-1.69	-0.93
R	-2.19	8.01	0.06	1.75	1.81	-4.75	-1.14	-0.53	-0.68	-6.07	-0.32	1.79	-2.56	-6.98	-5.42	-6.68	-6.17	-3.40	-4.42	-6.07
N	-0.09	0.06	1.02	2.59	-0.09	0.52	1.68	0.09	3.23	-1.19	-1.54	-1.31	-3.29	-3.07	-0.99	-0.50	1.27	-2.14	0.85	-3.64
D	-0.23	1.75	2.59	3.60	2.49	2.43	2.78	-2.49	1.47	-3.39	-5.87	-3.20	-3.88	-8.30	-2.10	0.18	1.56	-4.72	-3.74	-4.22
C	-0.03	1.81	-0.09	2.49	1.14	0.92	-0.26	0.49	1.81	1.07	-3.23	-1.00	-0.06	0.69	-2.93	0.46	-1.67	-0.91	-1.92	-0.41
Q	1.37	-4.75	0.52	2.43	0.92	2.18	2.07	0.58	-2.75	-2.96	-2.61	1.01	-3.45	-7.87	-1.67	-0.40	0.75	-4.29	-5.31	-2.96
E	0.43	-1.14	1.68	2.78	-0.26	2.07	2.47	-0.08	-1.59	-2.98	-3.46	0.52	-5.46	-7.88	-0.15	-1.09	0.84	-4.31	-1.71	-0.98
G	0.49	-0.53	0.09	-2.49	0.49	0.58	-0.08	4.86	-5.17	-0.47	-4.40	-0.03	-3.88	-5.13	-0.74	0.00	0.69	1.62	0.91	-2.75
H	-0.19	-0.68	3.23	1.47	1.81	-2.75	-1.59	-5.17	4.32	-0.07	-2.11	-1.88	-2.56	0.42	-5.42	0.49	0.48	-0.23	2.50	-2.07
I	-0.57	-6.07	-1.19	-3.39	1.07	-2.96	-2.98	-0.47	-0.07	1.97	0.55	-1.71	0.39	0.80	-3.64	-0.26	-1.22	1.78	1.36	3.22
L	-1.48	-0.32	-1.54	-5.87	-3.23	-2.61	-3.46	-4.40	-2.11	0.55	3.71	-0.95	1.91	1.37	-4.65	-2.05	-0.87	-5.27	0.88	0.63
K	1.00	1.79	-1.31	-3.20	-1.00	1.01	0.52	-0.03	-1.88	-1.71	-0.95	3.22	-4.20	-5.98	-2.42	-0.71	-2.20	-1.04	-1.14	-1.07
M	-2.59	-2.56	-3.29	-3.88	-0.06	-3.45	-5.46	-3.88	-2.56	0.39	1.91	-4.20	6.55	1.24	-0.96	0.61	-1.70	-2.11	1.52	2.39
F	-2.48	-6.98	-3.07	-8.30	0.69	-7.87	-7.88	-5.13	0.42	0.80	1.37	-5.98	1.24	5.70	-8.55	-6.64	-5.29	-6.53	0.79	-2.55
P	-0.81	-5.42	-0.99	-2.10	-2.93	-1.67	-0.15	-0.74	-5.42	-3.64	-4.65	-2.42	-0.96	-8.55	7.32	-1.61	-5.74	-4.97	-1.35	-1.30
S	0.41	-6.68	-0.50	0.18	0.46	-0.40	-1.09	0.00	0.49	-0.26	-2.05	-0.71	0.61	-6.64	-1.61	2.71	2.72	-4.23	-0.08	0.27
T	0.45	-6.17	1.27	1.56	-1.67	0.75	0.84	0.69	0.48	-1.22	-0.87	-2.20	-1.70	-5.29	-5.74	2.72	2.23	-5.72	-4.74	-1.22
W	0.42	-3.40	-2.14	-4.72	-0.91	-4.29	-4.31	1.62	-0.23	1.78	-5.27	-1.04	-2.11	-6.53	-4.97	-4.23	-5.72	8.71	0.03	-5.62
Y	-1.69	-4.42	0.85	-3.74	-1.92	-5.31	-1.71	0.91	2.50	1.36	0.88	-1.14	1.52	0.79	-1.35	-0.08	-4.74	0.03	4.30	-0.30
V	-0.93	-6.07	-3.64	-4.22	-0.41	-2.96	-0.98	-2.75	-2.07	3.22	0.63	-1.07	2.39	-2.55	-1.30	0.27	-1.22	-5.62	-0.30	3.92
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

Table 2: Twilight Matrix

A	-0.53	0.49	1.09	0.58	0.32	-0.37	-0.90	0.31	-3.06	-0.25	-0.38	-1.44	-0.46	0.18	-0.41	0.21	-0.43	0.44	0.01	0.39
R	0.49	0.40	-0.60	-0.04	-0.46	0.37	0.73	-0.01	-0.43	0.88	-0.35	-0.53	0.75	-0.36	-0.31	-0.29	0.39	0.28	-0.29	0.20
N	1.09	-0.60	1.19	0.52	-0.38	-1.14	-1.25	-0.38	0.37	0.20	-0.15	0.17	1.35	-0.84	-0.47	-0.11	0.07	-0.92	0.65	0.90
D	0.58	-0.04	0.52	-0.53	-0.29	-0.95	-0.48	-0.19	0.11	1.07	0.53	-0.47	-0.08	1.19	-0.99	-0.06	0.40	-0.73	0.54	0.39
C	0.32	-0.46	-0.38	-0.29	0.24	0.60	0.11	0.22	-1.65	-0.44	0.33	0.02	-0.54	-0.50	0.64	-0.46	-0.64	0.09	-0.33	-0.09
Q	-0.37	0.37	-1.14	-0.95	0.60	-1.59	-0.00	-0.42	1.61	-0.76	-1.18	0.58	-0.38	-0.77	0.65	-0.74	1.15	-3.68	-1.13	1.17
E	-0.90	0.73	-1.25	-0.48	0.11	-0.00	0.26	-0.27	-2.11	-0.14	0.41	0.33	-1.52	-0.23	0.19	-0.15	-0.38	1.13	1.40	-0.42
G	0.31	-0.01	-0.38	-0.19	0.22	-0.42	-0.27	0.02	1.35	0.20	0.15	-0.18	-0.37	0.63	0.04	0.47	-0.54	0.24	-0.63	-0.67
H	-3.06	-0.43	0.37	0.11	-1.65	1.61	-2.11	1.35	2.87	0.50	0.02	0.95	1.85	-0.54	-2.04	0.58	0.21	-1.45	0.13	-0.60
I	-0.25	0.88	0.20	1.07	-0.44	-0.76	-0.14	0.20	0.50	-1.87	0.38	0.30	1.18	-0.46	-1.49	0.02	-0.68	0.83	-0.55	0.56
L	-0.38	-0.35	-0.15	0.53	0.33	-1.18	0.41	0.15	0.02	0.38	-0.63	0.12	0.31	0.09	0.05	0.16	-0.13	-0.62	-0.16	-1.07
K	-1.44	-0.53	0.17	-0.47	0.02	0.58	0.33	-0.18	0.95	0.30	0.12	0.40	0.62	-0.99	0.25	0.79	-0.51	-1.65	0.33	-0.53
M	-0.46	0.75	1.35	-0.08	-0.54	-0.38	-1.52	-0.37	1.85	1.18	0.31	0.62	0.83	-0.73	-1.45	0.73	-1.98	0.70	-0.37	0.65
F	0.18	-0.36	-0.84	1.19	-0.50	-0.77	-0.23	0.63	-0.54	-0.46	0.09	-0.99	-0.73	-0.99	-0.16	0.64	0.73	0.62	0.72	0.22
P	-0.41	-0.31	-0.47	-0.99	0.64	0.65	0.19	0.04	-2.04	-1.49	0.05	0.25	-1.45	-0.16	0.92	-0.29	0.68	0.28	-1.39	-0.87
S	0.21	-0.29	-0.11	-0.06	-0.46	-0.74	-0.15	0.47	0.58	0.02	0.16	0.79	0.73	0.64	-0.29	-0.29	-0.27	-0.46	-0.17	-0.11
T	-0.43	0.39	0.07	0.40	-0.64	1.15	-0.38	-0.54	0.21	-0.68	-0.13	-0.51	-1.98	0.73	0.68	-0.27	1.15	-1.46	0.30	0.97
W	0.44	0.28	-0.92	-0.73	0.09	-3.68	1.13	0.24	-1.45	0.83	-0.62	-1.65	0.70	0.62	0.28	-0.46	-1.46	-1.12	1.42	0.25
Y	0.01	-0.29	0.65	0.54	-0.33	-1.13	1.40	-0.63	0.13	-0.55	-0.16	0.33	-0.37	0.72	-1.39	-0.17	0.30	1.42	0.14	0.01
V	0.39	0.20	0.90	0.39	-0.09	1.17	-0.42	-0.67	-0.60	0.56	-1.07	-0.53	0.65	0.22	-0.87	-0.11	0.97	0.25	0.01	0.07
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V

the diagonal of each matrix, indicating a high reward for keeping a certain residue.

A key difference between the matrices is the higher variation of numbers in the super family matrix. Due to the higher amount of evolutionary connection between the strings, there were more very high scores and very low scores when counting the pairs. This of course translates to more extreme scores in the resulting matrix. Compare this with the twilight matrix, which does have as many extreme scores. This is due to the fact that the strings are barely related if at all, making their comparisons and distribution of pairs much closer to random than the superfamily dataset.

0.4 Comparison to BLOSUM Matrices

The superfamily matrix appears closest to BLOSUM50. This is unsurprising as BLOSUM50 is creates using blocks with at most 50% homology. We know from the manual of the dataset that the superfamily groups have between 25%-50% homology, which places it squarely into the types of blocks that would be used to create BLOSUM50.

The twilight matrix instead appears more similar to BLOSUM42. 42 is the lowest benchmark used in a BLOSUM matrix, so this is again unsurprising. The main difference that allows me to make the assessment is that BLOSUM42 has a lot more -1 and 0 entries. This is much more similar to our twilight matrix, again with the more evenly distributed scores.

0.5 Conclusion

In summary, we were able to implement the matrix construction as seen in BLOSUM. We then took the created matrices and were able to see some qualitative similarities and differences, as well as try to line them up against some of the canonical BLOSUM matrices. Specifically BLOSUM42 and BLOSUM50 present several similarities to our matrices, allowing us to confirm our implementation.